Stanford Machine Learning W6 Advice.

1. Deciding What To do Next.

- Debugging → error ↑
  - larger training data.
  - smaller set of feature.
  - get additional features.
  - polynomial features.
  - ↑ λ or ↓ λ

  } how to pick one of the options?

- Machine Learning Diagnostic.

2. Evaluating a Hypothesis

(1) split data set (randomly sorted) → Training set 0.7, Test set 0.3

m_{test} = # of test examples,

(2) Procedure

(1) learn θ s.t. $\min_{\theta} J(\theta)$

(2) lin. reg. $J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} [h_\theta(x^{(i)}_{test}) - y^{(i)}_{test}]^2$ (linear reg.)

log. reg $J_{test}(\theta) = \frac{-1}{m_{test}} \sum_{i=1}^{m_{test}} y^{(i)}_{test} \cdot \log h_\theta(x^{(i)}_{test}) + (1-y^{(i)}_{test}) \log h_\theta(x^{(i)}_{test})$

misclassification error.

$$err(h_\theta(x), y) = \begin{cases} 1 & (h_\theta(x) \geq 0.5 \ \& \ y=0) \ \| \ (h_\theta(x) \leq 0.5 \ \& \ y=1) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Test error} = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_\theta(x^{(i)}_{test}), y^{(i)}_{test})$$

3. Model Selection and Train / Validation / Test set

- Overfitting : cost < generalization error.

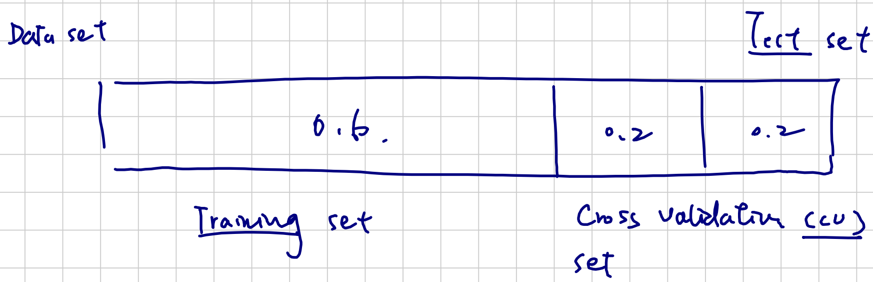
- Model selection (d = $\deg_x (h)$.)

$$
\begin{array}{llll}
d=1 & h = \theta_0 + \theta_1 x & \to \theta^{(1)} & \to J_{test}(\theta^{(1)}) \\
d=2 & h = \theta_0 + \theta_1 x + \theta_2 x^2 & \to \theta^{(2)} & \vdots \\
\vdots & \quad \vdots & & \\
d=10 & h = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}. & \to \theta^{(10)} & J_{test}(\theta^{(10)})
\end{array}
$$

- problem. $J_{test}(\theta)$ is likely to be an optimistic generalization error.

  (our extra parameter d is fit to test)

↓



Train / validation / test error.

Training error.
$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} [h_\theta(x^{(i)}) - y^{(i)}]^2$$

Cross validation error.
$$J_{train}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} [h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)}]^2$$

Test error
$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} [h_\theta(x_{test}^{(i)}) - y_{test}^{(i)}]^2$$

Procedure

get $\theta^{(d)}$ → test on cv set. $J_{cv}(\theta^{(d)})$ $\forall d$.

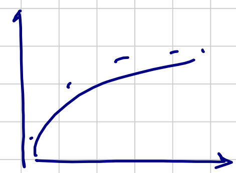
→ find d s.t. $J_{cv}(\theta^{(d)})$ is min.

→ estimate generalization error for test set $J_{test}(\theta^{(4)})$

4. Diagnosing Bias v.s. Variance.

high bias

underfit



high variance

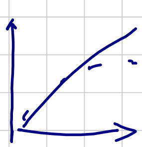
overfit

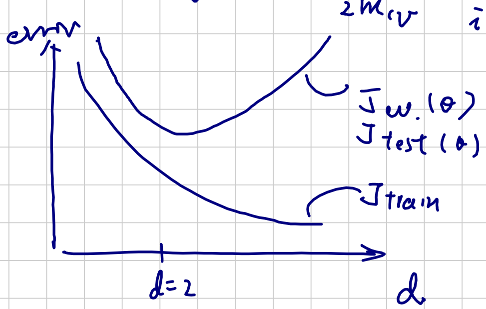
Bias / Variance

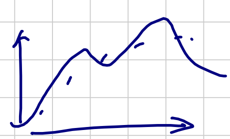
$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}[h_\theta(x^{(i)}) - y^{(i)}]^2$$

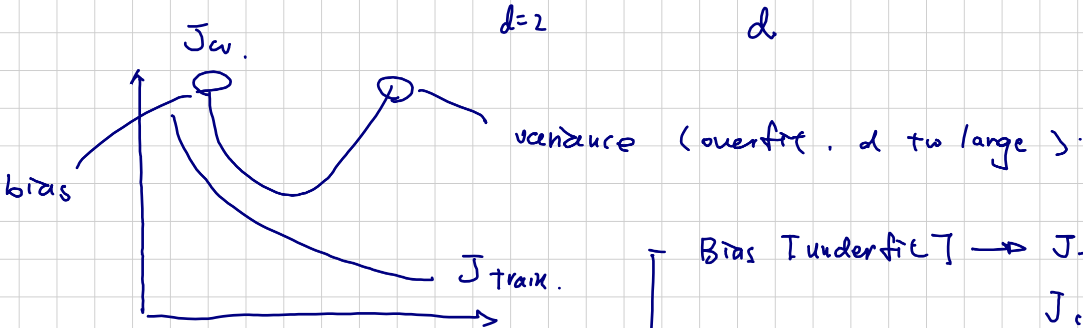
(or $J_{test}(\theta)$)

Cross validation error $J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}[h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)}]^2$









Bias [underfit] → $J_{train}(\theta)$ high
$J_{cv}(\theta) \approx J_{train}(\theta)$

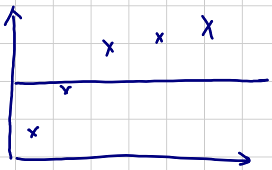
Variance [overfit] → $J_{train}(\theta)$ low
$J_{cv}(\theta) >> J_{train}(\theta)$

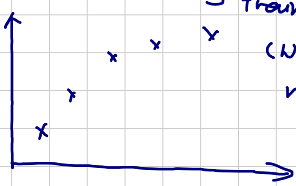
5. Regularization and Bias / Variance.

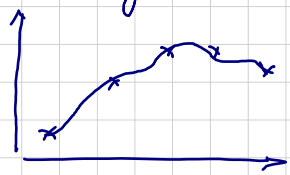
Linear regression with regularization.

Model: $h_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_4 x^4$

$$J(\theta) = \underbrace{\frac{1}{2m}\sum_{i=1}^{m} [h_\theta(x^{(i)}) - y^{(i)}]^2}_{J_{train}\ (w/o.\ reg)} + \underbrace{\frac{\lambda}{2m}\sum_{j=1}^{m} \theta_j^2}_{reg.}$$







large λ.

$\lambda = 10000$, $\theta_{1,2,3,4} \approx 0$.

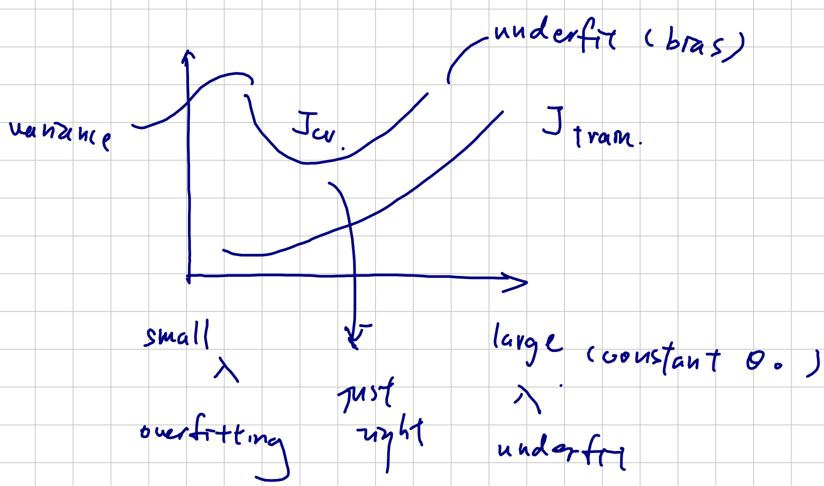
$h_\theta \approx \theta_0$.

high bias (underfit)

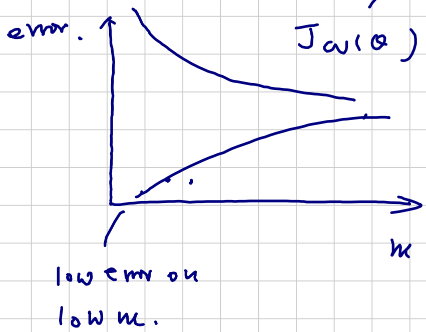
$\lambda = 0$ (small λ)

high variance (overfit)

Training error (J_{train}) vs. validation error (J_{cv})



6. Learning Curve.



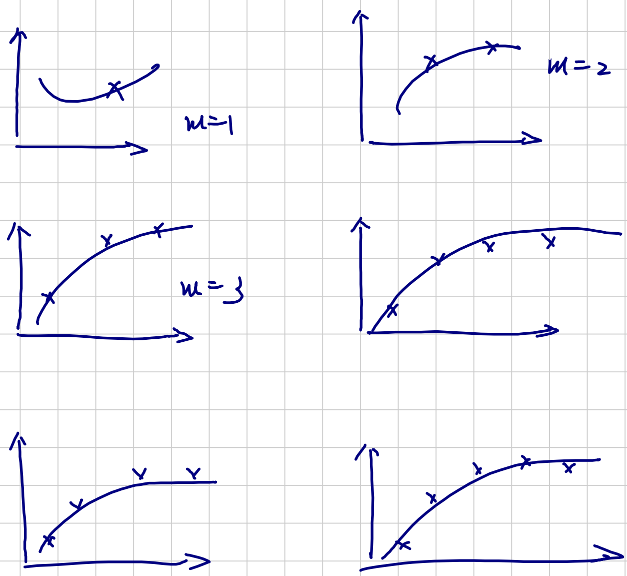
the more data, the better our model generalizes.

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left[h_\theta(x^{(i)}) - y^{(i)}\right]^2$$

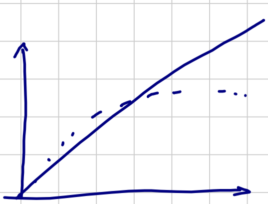
$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}\left[h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)}\right]^2$$

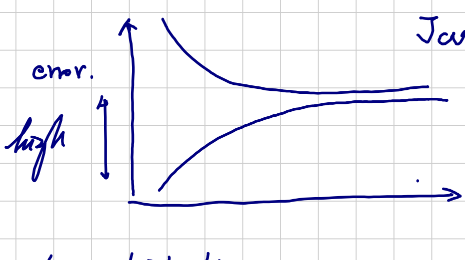
$\Rightarrow$ error $\sim$ m. "easier to fit in smaller training set"

for example $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$



High Bias



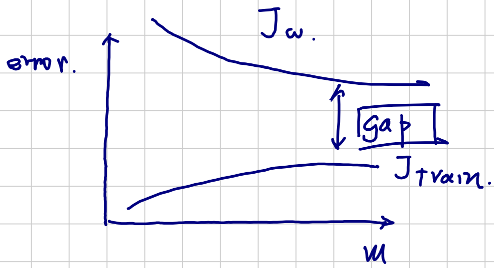


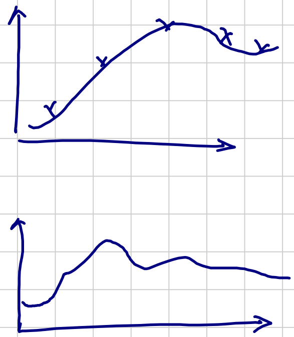
$J_{cv}(\theta)$ plateaus out.

$J_{train}(\theta) \approx J_{cv}$. (small data set)

If learning algorithm has high bias, ↑ data set does not help.

High Variance





deg h ↑
small λ.

If a learning algorithm is suffering from high variance, getting more training data is likely to help.

7. Deciding What to Do Next Revisited.

- get more examples : fix high variance.
- try smaller set of feature : fix high variance

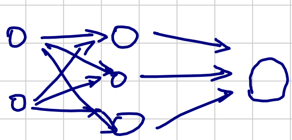
> over fitting.

- Add features : fix high bias.
- polynomial features : fix high bias

- try decr. λ ⟶ fix high bias (incr. importance of θ).
- try incr λ ⟶ fix high variance (decr. importance of θ).

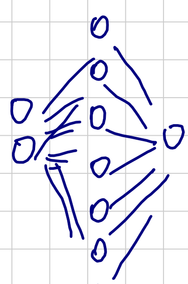
Neural Network and Overfitting

"Small" Neural network
(few parameters ⟶ prone to underfitting)



computationally cheap.

"large" Neural Network
(more parameters ⟶ prone to overfitting)



how many hidden layers ?

Computationally expensive.
⟶ use regularization (λ) to address overfitting.